Streamlining Literature Reviews Using an Automatic and Flexible Data Gathering and Classification Platform

António Miguel Martins
INESC-ID, Instituto Superior Técnico, Universidade de Lisboa
Lisboa, Portugal antonio.valente.martins@tecnico.ulisboa.pt

Alberto Rodrigues da Silva
INESC-ID, Instituto Superior Técnico, Universidade de Lisboa
Lisboa, Portugal alberto.silva@tecnico.ulisboa.pt

Jacinto Estima
INESC-ID, Lisboa, Portugal
CISUC, Dep. of Informatics Engineering, University of Coimbra
Coimbra, Portugal estima@dei.uc.pt

Abstract

Literature reviews are a crucial but time-consuming and complex task in scientific research. As such, interest in automating this process using machine learning techniques has increased over the last few years. In this paper, we present a method of streamlining the process of writing literature reviews by automating several aspects of the process using Maestro v2023, an automatic and flexible data gathering and classification platform. Maestro v2023 is a revamped version of the original Maestro platform, designed to be modular and configurable, allowing users in an organization to create search contexts that automatically gather and classify data for them. We analyze the work related to literature review automation and suggest how Maestro can contribute to this field, demonstrating how the system was utilized in order to streamline our own literature review process, as well aid us in formulating the abstract and extracting relevant keywords to this paper.

Keywords: Machine learning; Data classification; Data gathering; Automatic literature review; Text generation.

1. Introduction

Literature reviews are crucial in scientific research, gathering and consolidating current research to provide a comprehensive overview of knowledge in the field. However, writing literature reviews can be daunting, especially for less experienced members, such as postgraduate students [1]. As such, research on automating literature reviews has gained attention in recent years. Research on this topic was further propelled by the field of Machine learning (ML), which has harnessed the ability to access the vast amounts of digitized data available on the web, leading to the exponential growth and rapid improvement in major tasks.

Maestro was developed to streamline some of these processes, serving as a modular, extensible, and configurable platform for data gathering and data classification [2, 3]. Although the platform's primary goal is data gathering and classification, it also allows for additional modular steps, such as data filtering and post-processing, further expanding its range of applications. This paper aims to present how Maestro v2023, an improved version of the original platform, can be helpful to the scientific community, namely by automating several aspects of the literature review process. From now on, we will refer to the original Maestro platform as "Maestro v2022" and the proposed revamped version resulting from the expansion as "Maestro v2023". The term "Maestro" will refer to the overall concept of the platform.

2. Background

This section clarifies various literature review aspects and techniques used when automating this process.

2.1. Literature Reviews

Writing a literature review is often necessary when conducting research in a given scientific field or topic. It is often essential to demonstrate a certain degree of understanding of the field and bridge the gaps in the researcher's knowledge. Literature reviews differ from general exposition and contextualization on a given topic, as researchers are expected to provide critical evaluations and conclusions of both their given field of study, as well as the works inserted in such context, ideally demonstrating the motivation and arguments for the pursuit and value of their own research [4].

To conduct an effective literature review, researchers shall follow some general guidelines. A good example of a set of general guidelines is presented by the Royal Literary Fund [5], stating that researchers ought to survey the literature on the subject, present it in an organized fashion, synthesize it, and critically analyze it, thus demonstrating a familiarity with the subject by presenting and analyzing the current gaps, limitations, and controversies in the field.

Despite the sheer importance of this process, literature reviews are often dense and cumbersome, requiring a great effort to be done accurately. As such, methods to aid researchers in this process are constantly being developed, aiming to maximize the quality of the review while diminishing the time required to complete it. Nonetheless, and despite the potential decrease in the effort needed to conduct literature reviews, researchers must keep in mind the ethical concerns underlying the execution of this task, since an inaccurate, incomplete, or biased analysis could result in an incorrect understanding of context, as well as the unintended extrapolation of findings, by the reader [6].

2.2. Text Summarization

The area of text summarization aims to allow the condensation of documents and publications. When done correctly, the produced summaries are expected to highlight the critical aspects of these artifacts, effectively undermining the need to sift through a large amount of redundant information.

Different trends and techniques form the basis for research within this field. A recent study by Widyassari et al. [7] systematically reviews automatic text summarization by analyzing different publications published from 2008 to 2019. They identified ML approaches as the most predominant technique, being used in more than half of the studies analyzed. Regarding trends, multi-document summarization was the most prevalent, in which the summary is generated based on a set of input documents and the target is to remove repetitive content in the input documents [8]. Extractive summarization followed closely behind, an approach focused on choosing the most important words, sentences, and paragraphs to produce a summary. The third most common trend was abstractive summarization, which aims to produce summaries consisting of sentences different from the original document(s). Abstractive approaches tend not to be as favored in research as extractive approaches, as they are highly complex and require extensive natural language processing (NLP) [8].

2.3. Text Simplification

Text simplification aims to make complex language easier to understand by rephrasing it into simpler terms and typically involves making use of three core elements: splitting, deletion, and paraphrasing. Splitting involves breaking lengthy sentences into several smaller sentences that enhance the readability of the overall text. Deletion discards a sentence's extraneous and less consequential parts, thereby reducing its complexity. Finally, paraphrasing is used to reorder, substitute, and, in some cases, expand sentence constructs to achieve a simplified version of the original text [9].

Research in this area has various practical applications, including assisting people with disabilities, low literacy, non-native language backgrounds, or limited expertise to comprehend written materials more easily [9]. Despite its complexity, the automation of this process has rapidly grown, spurred by the rise of both ML and NLP [10].

3. Evolution of the Maestro Platform

This section presents the underlying philosophy and architecture of Maestro, along with a timeline of its development.

3.1. Maestro's Architecture

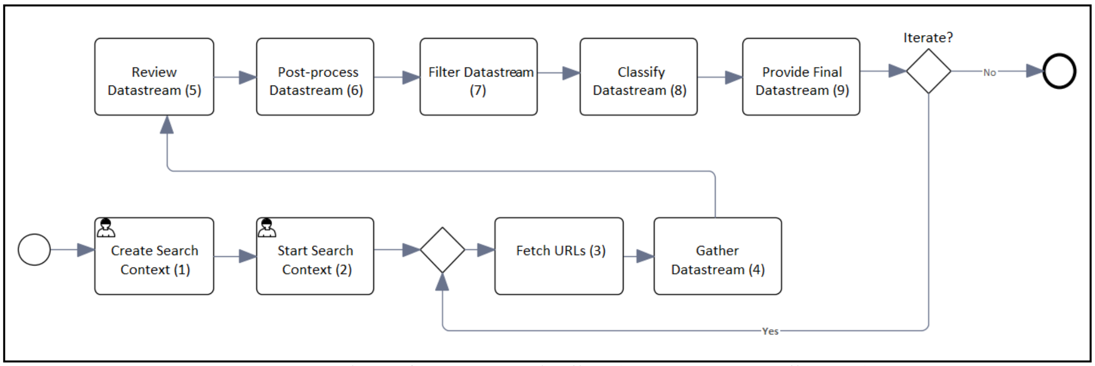


Fig. 1. Overview of Maestro's pipeline (BPMN process diagram).

Maestro was created to gather and classify data as a service. It functions in a modular, extensible, and configurable fashion, enabling users within an organization to automatically collect and classify data of various types (e.g., images, sound, text).

Maestro possesses three key concepts: organizations, users, and search contexts. Maestro users can be associated with one or more organizations, which exist to facilitate collaboration and simultaneous workflows. Users can define multiple workflows configured to gather, classify, and deliver their target data. These workflows are named search contexts, and function as declarative expressions of the tasks to be run through Maestro's pipeline.

Plugins-based Pipeline

As illustrated in Fig. 1, Maestro supports a pipeline that, once run, results in a classified dataset that can be provided to external services. Maestro's pipeline comprises nine essential steps or phases, namely: (1) Create / Configure a search context; (2) Start a configured search context; (3) Fetch URLs pointing to objects of the desired data type; (4) Gather the resources or data items from the fetched URLs; (5) Review the gathered data items and manually discard those deemed irrelevant; (6) Post-process the gathered data with the use of plugins, acquiring additional parameters for the subsequent steps (e.g., adding metadata to image data). (7) Filter the data according to the specified plugins and parameters defined by the user (e.g., filtering based on the date and location of a given data item). (8) Classify the dataset items using the desired classification plugins; (9) Provide the resulting classified dataset to external services.

Plugins are user-made scripts that follow a common interface that Maestro understands. Of these phases, the only two that cannot be configured to use plugins are the gathering (4) and providing (9) phases, though the latter can be configured to send the results to the desired endpoint.

3.2. Maestro's Iterations

The original iteration of Maestro, referred to as Maestro v2022 [2, 3], implemented the majority of systems detailed previously. It allows users to use the system to gather and classify both image data and sound data types.

However, Maestro v2022 still bore some limitations that needed to be addressed, namely: limited types of data, lack of automatability, inability to do in-depth analysis of the pipeline, and other constraints. Thus, a new iteration of the platform, Maestro v2023, began being developed to expand and refine Maestro's capabilities.

Maestro v2023 is currently in development. However, some of the previously mentioned limitations have already been addressed. Most importantly, the platform's expansion to allow the use of text data types in its pipeline and the ability to utilize multiple classification algorithms, are fundamental for the process of automating the literature review process. Furthermore, one significant change that has taken effect in Maestro v2023 is the deviation from merely classification tasks. This new iteration has an increasing focus on allowing users to perform other ML techniques during the "classification stage" of the pipeline (e.g., text summarization), significantly increasing the platform's utility.

Table 1. Description of the developed plugins for supporting scientific publications.

Name	Plugin Type	Description
Elsevier Fetcher	Fetcher	Queries the Elsevier API [14] for scientific publications using a search string and tags.
ArXiv Fetcher	Fetcher	Queries the ArXiv API [15] for scientific publications using a search string and tags.
Duplicate Filter	Filter	Removes duplicates of scientific publications by comparing the descriptor (title or DOI).
Paper Summarizer	Classifier	Generates summaries of scientific publications using the BARTxiv model [18].
Abstract Simplifier	Classifier	Rewrites difficult-to-understand scientific abstracts into simpler, easier-to-read versions using the SAS model [17].
Keyword Extractor	Classifier	Extracts relevant keywords from paper abstracts using KeyBERT [19].

4. Demonstration

To demonstrate Maestro's usefulness in the literature review process, we showcase part of the methodology behind the literature review process for this research, aided by the use of Maestro. To do this, our team has developed a set of plugins that, by making use of external tools, has allowed us to streamline the literature review process for this article. Table 1 describes the developed plugins. The following steps were followed to make use of Maestro, along with the developed plugins.

4.1. Essential Configurations

The user creates a search context through Maestro's interface. The user must define an owner, title, unique code, and a description for their search context.

Once created, the user must configure their search context. The user defines the essential configurations, which are mandatory. As illustrated in Fig. 3, he defines the search string for finding the data ("automatic literature review"), relevant tags ("system", "machine learning", "scientific paper"), the data type ("Text"), as well as other options that allow the search context to automatically run again after a certain amount of time (in this case, we set it to "Don't repeat").

4.2. Advanced Configurations

The user can then define the advanced configurations. Despite these settings not being mandatory, the system will do nothing if they are not configured. Generally, users may define settings for each phase performed automatically by Maestro's pipeline, fetching, gathering, post-processing, filtering, classifying, and providing the datastream(s). In this phase, the user shall proceed by conducting the following tasks:

Select the "Elsevier API" and "ArXiv API" fetching plugins for fetching URLs of publications related to the search string and tags; select the "Paper Summarizer", "Abstract Simplifier", "Keyword Extractor" plugins, to be considered during the classification step, as illustrated in Fig. 4; apply filtering configurations to discard any duplicates of gathered articles, by checking their DOI and/or Title; the user may also specify the configurations for an HTTP Rest endpoint to which the data will be sent during the providing step. For this scenario, no post-processing plugin was applied.

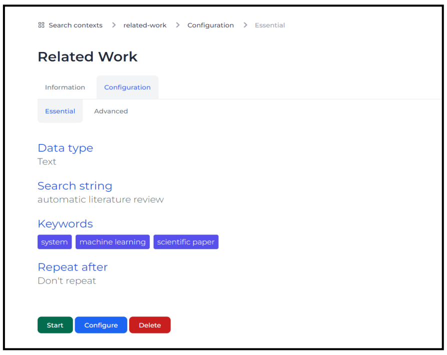
4.3. Search Context Run and Results

Once the configurations have been defined, the user triggers the run of the search context, and waits for the results. This process runs in the background, and may take some time to complete. As illustrated in Fig. 5, the system provided the user with multiple scientific publications related to the defined search strings and tags, summarized them, simplified the abstract, and extracted relevant keywords from the original abstract. Selecting the "Show Details" option allows users to see each data object in finer detail (See Fig. 6).

The configured search context was run two consecutive times, using similar configurations of tags: the first run used the tags "system", "machine learning" and "scientific paper"; the second run used the tags "text summarization", "machine learning" and "scientific paper". Both runs used the same search string.

The provided dataset includes 55 different publications after removing those flagged as duplicates. Out of the gathered publications, 45 were fetched using the "ArXiv Fetcher" plugin, with the remaining 10 being fetched by the "Elsevier Fetcher" plugin. Out of the 55 articles, we identified 8 as potentially relevant to our research, with 3 of them being integrated into the final version of this paper [7, 12, 13].

4.4. Additional Scenario: Abstract Formulation and Keyword Extraction



Related Work

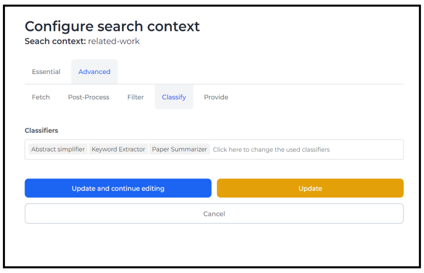
Data type
Text

Search string
automatic literature review

Keywords
system | machine learning | scientific paper

Repeat after
Don't repeat

Fig. 3. Configuration of search context in the Maestro platform.

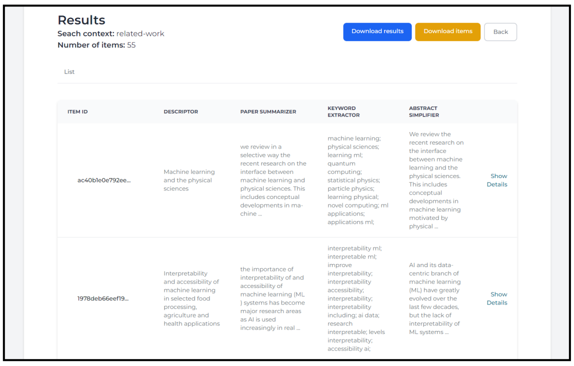

Essential
Advanced
Fetch
Post-Process
Filter
Classify
Provide
Click here to change the used classifiers
Update and continue editing
Update
Cancel

Configure search context

Seach context: related-work

Classifiers
Abstract simplifier | Keyword Extractor | Paper Summarizer

Fig. 4. Configuration of three data classification plugins for the classification stage of a search context: Paper Summarizer, Abstract Simplifier and Keyword Extractor.



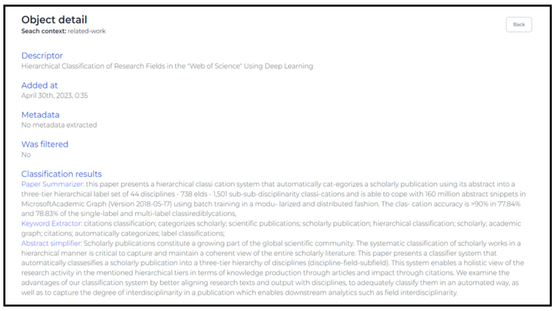
Results

Seach context: related-work

Number of items: 55

ITEM ID	DESCRIPTOR	PAPER SUMMARIZER	KEYWORD EXTRACTOR	ABSTRACT SIMPLIFIER	
ac40b1e0e792ee...	Machine learning and the physical sciences	we review in a selective way the recent research on the interface between machine learning and physical sciences. This includes conceptual developments in machine ...	machine learning; physical sciences; learning ml; quantum computing; statistical physics; particle physics; learning physical; novel computing; ml applications; applications ml;	We review the recent research on the interface between machine learning and the physical sciences. This includes conceptual developments in machine learning motivated by physical ...	Show Details
1978deb66eef19...	Interpretability and accessibility of machine learning in selected food processing, agriculture and health applications	the importance of interpretability of and accessibility of machine learning (ML) systems has become major research areas as AI is used increasingly in real ...	interpretability ml; interpretable ml; improve interpretability; interpretability accessibility; interpretability; interpretability including; ai data; research interpretable; levels interpretability; accessibility ai;	AI and its data-centric branch of machine learning (ML) have greatly evolved over the last few decades, but the lack of interpretability of ML systems ...	Show Details

Fig. 5. Resulting dataset from a search context configured to gather and classify scientific publications related to "Automatic Literature Review".



Object detail

Seach context: related-work

Descriptor
Hierarchical Classification of Research Fields in the "Web of Science" Using Deep Learning

Added at
April 30th, 2023, 0:35

Metadata
No metadata extracted

Was filtered
No

Classification results
Paper Summarizer: this paper presents a hierarchical classi cation system that automatically cat-egorizes a scholarly publication using its abstract into a three-tier hierarchical label set of 44 disciplines - 738 elds - 1,501 sub-sub-disciplinarity classi-cations and is able to cope with 160 million abstract snippets in MicrosoftAcademic Graph (Version 2018-05-17) using batch training in a modu- larized and distributed fashion. The clas- cation accuracy is >90% in 77.84% and 78.83% of the single-label and multi-label classirediblycations,
Keyword Extractor: citations classification; categorizes scholarly; scientific publications; scholarly publication; hierarchical classification; scholarly; academic graph; citations; automatically categorizes; label classifications;
Abstract simplifier: Scholarly publications constitute a growing part of the global scientific community. The systematic classification of scholarly works in a hierarchical manner is critical to capture and maintain a coherent view of the entire scholarly literature. This paper presents a classifier system that automatically classesifies a scholarly publication into a three-tier hierarchy of disciplines (discipline-field-subfield). This system enables a holistic view of the research activity in the mentioned hierarchical tiers in terms of knowledge production through articles and impact through citations. We examine the advantages of our classification system by better aligning research texts and output with disciplines, to adequately classify them in an automated way, as well as to capture the degree of interdisciplinarity in a publication which enables downstream analytics such as field interdisciplinarity.

Fig. 6. Detailed view of one of the provided data objects' properties.

Maestro also allows users to manually submit the data to be used in Maestro's pipeline, rather than having this data be obtained during the fetching and gathering stages of the pipeline.

We used this feature to run a draft of this paper through the pipeline, to summarize it using the "Paper Summarizer" plugin. This summarization was then inserted into a separate run of our pipeline and, using the "Abstract Simplifier" plugin, produced a more coherent version of this summarization. Finally, after some modification to the produced artifact, we passed the result to Maestro once again and used the "Keyword Extractor" plugin to identify potential keywords to be used in the final version of the paper (For an overview of the details and results, see https://tinyurl.com/45v2xruf).

5. Literature Review

This section presents work related to the field of automating literature reviews, written with the assistance of Maestro, a critical analysis and discussion of these works, as well as a detail of Maestro's contributions to the field considering this analysis.

5.1. Related Work

Much research has been conducted on automating and accelerating the literature review process using of ML and NLP models. Classification of scientific papers, scientific research summarization, and APIs for the retrieval of publications from repositories are just some of the many possible tools that enable researchers to streamline their literature review processes.

Bacinger et al. [11] designed a system aimed at semi-automating literature reviews. The system enables its users to query numerous sources for scientific papers. The users may then manually mark a subset of the provided papers as either positive or negative, which enables the training of a model to automatically classify the remaining papers, resulting in a finalized dataset of papers deemed relevant for the literature review.

Yuan et al. [12] present a model that summarizes and creates reviews for scientific articles. Though the results suggest the system cannot fully automate the scientific review process, it presents promising metrics demonstrating the possibility of streamlining this process when used in tandem with human reviewers. Wang et al. [13] presents ReviewRobot, which automatically assigns a review score to a given paper, and writes comments for multiple categories, such as novelty and clarity of the paper. Using their system, researchers can quickly identify the pros and cons of publications and more efficiently and accurately perform literature reviews.

Finally, the backbone of many of these works is the external access to various sources and repositories of scientific literature. The use of APIs, such as the one provided by Elsevier [14] and ArXiv [15], enable the query of their system for publications, enabling the development of external tools and solutions for discovering and retrieving said artifacts.

A recent study [16] provides an analysis of the current state of the art in automating literature review. The authors concluded that currently no system enables the full automation of literature review across multiple disciplines or even presents compatibility toward various sources of scientific publications. They describe that the most successful approaches tend to be semi-autonomous systems in which part of the literature review process is automatic and the remaining is manual.

5.2. Critical Analysis

Despite the steady advancement in this field, it is not possible to fully automate literature reviews. Most of the work done in this area tends to be streamlining this process by automating specific aspects, with the remainder done manually.

While much of the work done in this field is promising, aiding researchers in identifying gaps in their knowledge of a given subject and allowing them to explore these topics from different perspectives, they are often held back by several constraints. Many of the existing systems are limited either in domain or techniques used. Furthermore, while much of this research tends to innovate in several aspects, bridging the gap between these works is still lacking, preventing researchers from simultaneously using different sets of tools. Furthermore, independently of the quality of the techniques developed and the output from literature review automation tools, a certain degree of bias will always be present, which could exacerbate ethical concerns or even result in inaccurate literature reviews. However, this problem existed even before the development of these systems and should not prevent continued research into this subject. Nonetheless, researchers should be aware of this when using these tools and rely on their own experience to mitigate these concerns.

5.3. Maestro's Contributions

Considering our analysis of the field of literature review automation, we believe our approach can contribute to the field. While the strategy employed in Section 4 presents only a fraction of the possible paths that researchers may take when using Maestro to tackle the automation of literature reviews, we believe it to be a promising approach, as it allows researchers to agglomerate many different semi-automated approaches into a single system, maximizing the automatability of the process and potential of the field. Furthermore, by implementing this approach in a flexible system like Maestro, novel research can be added posteriorly, further increasing the quality of the process.

However, the system is not without its flaws. While it is modular, certain restrictions are still necessary to ensure the platform's functionality. The size of the datasets provided must be, at times, throttled in order not to overwhelm the system. Furthermore, while the platform can be used "out of the box" for many scenarios, developing plugins is still necessary to adapt Maestro to specific requirements, which can be complex, depending on the task. Individual plugins are also subject to the biases and restraints of the tools and services used in their development. One possible solution to this problem could be to apply a wide set of plugins to diminish the level of bias in the results.

6. Conclusion

Although literature reviews are crucial to scientific research, they can be time-consuming and complex. Consequently, researchers have been exploring ways to automate this process. The application of ML techniques has gained significant attention in recent years, given its tremendous growth and potential to advance the field further.

Currently, it is not yet possible to fully automate this process, with all available solutions being limited to some capacity. As such, we demonstrate how the use of Maestro v2023, a flexible platform to automatically gather and classify data, can be used as a means to aggregate many of the currently available tools for automating literature reviews. By making use of this approach, we were able to gather different publications on the topic of automatic literature reviews and, with the development of plugins that leveraged several ML techniques, were able to easily identify those relevant to our own literature review process. Furthermore, through the help of these plugins, a draft of the abstract for this paper was generated, along with a set of relevant keywords.

Despite its limitations, we consider this system to offer an advantage over existing works in the field, as it enables researchers to unify diverse techniques into a cohesive pipeline, bridging the gap between them. Additionally, it allows for iterative enhancements by incorporating novel solutions, further improving the overall process.

Acknowledgments

Research partially funded by FCT 02/SAICT/2017/29360, UIDB/50021/2020, and UIDB/00326/2020.

References

1. Shahsavar, Z., & Kourepaz, H. M. (2020). Postgraduate students' difficulties in writing their theses literature review. Cogent Education, 7(1). https://doi.org/10.1080/2331186x.2020.1784620
2. Serra, Alexandre & Estima, Jacinto & Rodrigues da Silva, Alberto. (2022). Maestro: An Extensible General-Purpose Data Gathering and Classification System. Proceedings of ISD'2022. A15. 10.13140/RG.2.2.26824.80646.

3. Magalhães Serra, A., Estima, J., & Rodrigues da Silva, A. (2023). Evaluation of Maestro, an extensible general-purpose data gathering and data classification platform. Information Processing & Management [*in Press*].
4. Literature review. (2022, August 29). The University of Edinburgh. https://www.ed.ac.uk/institute-academic-development/study-hub/learning-resources/literature-review
5. Royal Literary Fund. (2014, August 7). What is a literature review? - The Royal Literary Fund. https://www.rlf.org.uk/resources/what-is-a-literature-review/
6. Suri, H. (2020). Ethical Considerations of Conducting Systematic Reviews in Educational Research. In O. Zawacki-Richter, M. Kerres, S. Bedenlier, M. Bond, & K. Buntins (Eds.), Systematic Reviews in Educational Research (pp. 41–54). Springer Nature. https://doi.org/10.1007/978-3-658-27602-7_3
7. Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A., & Setiadi, D. R. I. M. (2020). Review of automatic text summarization techniques & methods. Journal of King Saud University - Computer and Information Sciences, 34(4), 1029–1046. https://doi.org/10.1016/j.jksuci.2020.05.006
8. El-Kassas, W. S., Salama, C., Rafea, A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 165, 113679. https://doi.org/10.1016/j.eswa.2020.113679
9. Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in Current Text Simplification Research: New Data Can Help. Transactions of the Association for Computational Linguistics, 3, 283–297. https://doi.org/10.1162/tacl_a_00139
10. Al-Thanyyan, S., & Azmi, A. M. (2021). Automated Text Simplification. ACM Computing Surveys, 54(2), 1–36. https://doi.org/10.1145/3442695
11. Bacinger, F., Boticki, I., & Mlinarić, D. (2022). System for Semi-Automated Literature Review Based on Machine Learning. Electronics, 11(24), 4124. https://doi.org/10.3390/electronics11244124
12. Yuan, W., Liu, P., & Neubig, G. (2021). Can We Automate Scientific Reviewing? Journal of Artificial Intelligence Research, 75, 171–212. https://doi.org/10.1613/jair.1.12862
13. Wang, Q., Zeng, Q., Huang, L., Knight, K., Ji, H., & Rajani, N. (2020). ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis. Proceedings of the 13th International Conference on Natural Language Generation, 384–397.
14. Elsevier Developer Portal. (n.d.). Elsevier. https://dev.elsevier.com/
15. ArXiv API Access - arXiv info. (n.d.). ArXiv. https://info.arxiv.org/help/api/index.html
16. Tsunoda, D. F., Da Conceição Moreira, P. S., & Guimarães, A. L. S. (2020). Machine learning e revisão sistemática de literatura automatizada: uma revisão sistemática. Revista Tecnologia E Sociedade. https://doi.org/10.3895/rts.v16n45.12119
17. Wang, H. (2022). Scientific abstract simplification. Hugging Face. https://huggingface.co/haining/scientific_abstract_simplification
18. Du, J. (2022). BARTxiv. Hugging Face. https://huggingface.co/kworts/BARTxiv
19. Grootendorst, M. (2022). KeyBERT: Minimal keyword extraction with BERT. GitHub. https://github.com/MaartenGr/KeyBERT